Hierarchical clustering for the identification of communities in networks

Sonia Cafieri, Pierre Hansen, Leo Liberti

Hierarchical clustering for the identification of communities in networks

Sonia Cafieri[1], Pierre Hansen[2,3], Leo Liberti[3]

[1] Dept. Mathématiques et Informatique, ENAC, 7 av. E. Belin, 31055 Toulouse, France
sonia.cafieri@enac.fr
[2] GERAD, HEC Montréal, Canada
pierre.hansen@gerad.ca
[3] LIX, École Polytechnique, 91128 Palaiseau, France
liberti@lix.polytechnique.fr

Mots-clés : *clustering, network, graph, community, modularity, hierarchical heuristic*

1 Introduction

The analysis of networks and in particular the identification of *communities*, or *clusters*, is a topic of active research and attracts an increasing attention in the operations research as well as the physics communities. Complex systems arising in a variety of fields can be represented as networks, or graphs, where the set of vertices is given by the entities under study and the edges represent relations holding for pairs of vertices. A typical example is given by social networks, modeling interactions among people. Other real-life applications include communicatons networks, such as the World Wide Web, and transportation networks, representing movements of people or goods.

Given a clustering criterion, the problem of community detection in networks can be formulated as an optimization problem. The current mainstream approach is based on the criterion proposed by Newman and Girvan in [1]. These authors introduced the concept of *modularity* for a partition of a network, defined as the sum for all communities of the difference between the fraction of edges they contain and the expected fraction of edges they would contain if all edges were drawn at random, keeping the same degree distribution. Maximizing modularity gives an optimal partition with its optimal number of clusters.

Community detection based on modularity maximization is currently done with hierarchical as well as with partitioning heuristics, hybrids and, in a few papers, exact algorithms. See [2] for an in-depth survey and [3] for recently proposed exact algorithms. Partitioning schemes aim at finding a single partition or possibly several partitions into given numbers of clusters. They are based on a variety of approaches, including genetic search and simulated annealing. Hierarchical heuristics aim at finding a set of nested partitions. They are in principle devised for finding a hierarchy of partitions implicit in the given network when it corresponds to some situation where hierarchy is observed or postulated. This is often the case, for instance, in social organization and evolutionary processes. Hierarchical heuristics can be further divided into agglomerative and divisive ones. Given a graph $G = (V,E)$ with $|V| = n$, agglomerative heuristics proceed from an initial partition with n communities each containing a single entity and iteratively merge the pair of entities for which merging increases most the objective function (e.g., modularity). Divisive heuristics proceed from an initial partition containing all entities and iteratively divide a community into two in such a way that the increase in the objective function value is the largest possible, or the decrease in the objective value is the smallest possible. Mergings or bipartitions can be ended once they do not improve the objective function value anymore. Results can be presented on a dendrogram, which displays mergings or divisions of communities.

2 A new locally optimal divisive heuristic

In this work, we consider the case of hierarchical networks and propose a divisive heuristic which is locally optimal, in the sense that each of the successive bipartitions is done in a provably optimal way. The bipartition subproblem is expressed as a quadratic mixed-integer program with a convex relaxation. To that effect, first we write the modularity Q as a function, for each community, of its number of inner edges and of the sum of degrees of its vertices :

$$Q = \sum_s \left[\frac{m_s}{m} - \left(\frac{d_s}{2m} \right)^2 \right], \tag{1}$$

where $m = |E|$ and m_s and d_s denote respectively the number of edges and the sum of degrees of the vertices of community s. Here $s \in \{1,2\}$ since we aim to find a bipartition. We express the sum of degrees d_2 of vertices belonging to the second community as a function of the sum of degrees d_1 of vertices belonging to the first one : $d_2 = d_t - d_1$, where d_t is the sum of degrees in the community to be bipartitioned. Hence, Q can be rewritten as :

$$Q = \frac{m_1 + m_2}{m} - \frac{d_1^2}{4m^2} - \frac{d_2^2}{4m^2} = \frac{m_1 + m_2}{m} - \frac{d_1^2}{2m^2} - \frac{d_t^2}{4m^2} + \frac{d_t d_1}{2m^2}. \tag{2}$$

We introduce binary variables X_{r1}, X_{r2} and Y_{i1} to identify to which community each vertex and each edge belongs. For $r = 1,2,\dots m$ and $s = 1,2$, X_{rs} is equal to 1 if edge r belongs to community s and 0 otherwise, and for $i = 1,2,\dots n$ Y_{i1} is equal to 1 if vertex i belongs to the first of the two communities of the bipartition. We then impose constraints to ensure consistency, i.e. that any edge $r = \{v_i, v_j\}$ with end vertices indiced by i and j can only belong to community s if both of its end vertices belong also to that community, and express m_s and d_1 in terms of the considered variables : $m_s = \sum_r X_{rs}$, $d_1 = \sum_{i \in V_1} k_i Y_{i1}$. We obtain a quadratic convex mixed-integer program which can be solved by `CPLEX`.

We use this bipartitioning method for the splitting step in a hierarchical divisive scheme. Hence, our divisive heuristic is based on bipartitions which are done at each step in an optimal way. We present a comparison with the spectral-based hierarchical divisive heuristic of Newman [4] and with the hierarchical agglomerative heuristic of Clauset et al. [5] and we show that the proposed locally-optimal divisive heuristic gives better results.

Références

[1] M. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69 :026133, 2004.

[2] S. Fortunato. Community detection in graphs. *Physics Reports*, 486 :75-174, 2010.

[3] D. Aloise, S. Cafieri, G. Caporossi, P. Hansen, L. Liberti, S. Perron. Column generation algorithms for exact modularity maximization in networks. *Physical Review E*, 82(4) :046112, 2010.

[4] M. Newman. Modularity and Community Structure in Networks. *Proceedings of the National Academy of Sciences, USA*, 103(23) :8577-8582, 2006.

[5] A. Clauset, M. Newman, C. Moore. Finding community structure in very large networks. *Physical Review E*, 70 :066111, 2004.